DOCUMENT RESUME

ED 279 711 TM 870 126

AUTHOR Boldt, Robert F.
TITLE Generalization of SAT Validity across Colleges.
INSTITUTION College Entrance Examination Board, New York, N.Y.
REPORT NO CEEB-RR-86-3; ETS-RR-86-24
PUB DATE 86
NOTE 17p.
AVAILABLE FROM College Board Publications, Box 886, New York, New York 10101 ($6.00).
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01 Plus Postage. PC Not Available from EDRS.
DESCRIPTORS Class Rank; College Admission; *College Entrance Examinations; Generalization; Grade Point Average; Higher Education; Mathematics Tests; *Predictive Validity; *Predictor Variables; Standardized Tests; *Test Theory; *Test Validity; Verbal Tests
IDENTIFIERS College Board Validity Study Service; *Scholastic Aptitude Test

ABSTRACT

This study investigated whether or not the validity of the Scholastic Aptitude Test (SAT) is both higher and less variable across colleges than it appears to be. Data from 99 validity studies conducted by the Validity Study Service of the College Board were used. In addition to test validities based on first-year college averages, which were calculated using institutional data, validities for each college were also estimated for two other groups--applicants for admission to the colleges, and all SAT takers. These last two estimates were based on range restriction theory. The results revealed that the average validity of both the SAT-V and SAT-M was estimated to be higher both for all test takers and for groups of applicants than for test takers on whom validity studies were based. A hypothesis that validities for SAT-V and SAT-M might differ across institutions but have the same ratio was also tested. It was thought that departures from this second assumption might lead to the detection of institutional types. Significant departures from the equal-ratio hypothesis tested at the 5 percent level occurred for about 5 percent of the institutions, so no detection of institutional types occurred. (Author/JAZ)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED279711

College Board
Report
No. 86-3

Generalization of SAT® Validity Across Colleges

Robert F. Boldt

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

P. Hendel

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Generalization of SAT® Validity Across Colleges

Robert F. Boldt

College Board Report No. 86-3
ETS RR No. 86-24

College Entrance Examination Board, New York, 1986

Robert F. Boldt is a Senior Research Scientist at Educational Testing Service, Princeton, New Jersey.

Researchers are encouraged to express freely their professional judgment. Therefore, points of view or opinions stated in College Board Reports do not necessarily represent official College Board position or policy.

The College Board is a nonprofit membership organization that provides tests and other educational services for students, schools, and colleges. The membership is composed of more than 2,500 colleges, schools, school systems, and education associations. Representatives of the members serve on the Board of Trustees and advisory councils and committees that consider the programs of the College Board and participate in the determination of its policies and activities.

Additional copies of this report may be obtained from College Board Publications, Box 886, New York, New York 10101. The price is $6.

Copyright © 1986 by College Entrance Examination Board. All rights reserved.
College Board, Scholastic Aptitude Test, SAT, and the acorn logo are registered trademarks of the College Entrance Examination Board.
Printed in the United States of America.

CONTENTS

Executive Summary by Donald E. Powers i

Abstract 1

Introduction 2

Assumptions and Hypotheses 3

Procedures 4

Results 6

Discussion 7

References 7

Appendixes
A. Tables 9
B. Use of Test Theory to Represent the Effects of Self-Selection 11
C. Use of a Supplementary Variable When Data Are Missing for an Explicit Selector 11
D. Generalizing the Assumption that the Validities Are Proportional Across Institutions 12
E. Calculating Validities in the Restricted Group 12

Tables

1. Means, Standard Deviations, Reliabilities, Intercorrelations, and Number of Cases for SAT Administrations in 1979 and 1980 9

2. Mean Validities for the VSS Groups, Observed and Implied, Based on the Two-Generalization Hypotheses Applied in the Three Sets of Groups 9

3. Standard Deviations of Validities for the VSS Groups, Observed and Implied, Based on the Two-Generalization Hypotheses Applied in the Three Sets of Groups 10

4. VSS-Group Correlations Between the Observed and Implied Validities, Based on the Two-Generalization Hypotheses Applied in the Three Sets of Groups 10

5. Percentage of VSS-Group Validities Accounted for by Sampling Error and the Equal-Validity Hypothesis Applied in the Three Sets of Groups 10

6. Means, Standard Deviations, and Fifth-Percentile Values of Test and True Score Validities for the VSS, Applicant, and SAT-Taker Groups 10

EXECUTIVE SUMMARY

Traditionally, the validity of a test has been judged according to how well the test performs in a particular situation or institution, as evidenced by data collected solely in that situation without regard to data collected in other similar situations. This emphasis on the need for local validation was reinforced by the apparent variation in validity among situations or institutions. It became increasingly apparent, however, that much of the observed variation was a result of statistical artifacts, such as sampling error and differences among sites in the range of test scores and in the reliability of the criterion. Just how much of this apparent variation among sites could be explained (or was generalizable) by these purely statistical factors, and how much might result from more substantive differences among situations was open to empirical study. A number of empirical methods were developed and applied in numerous situations, mostly employment settings, and generally the findings have supported the hypothesis of validity generalization.

There has been less attention to validity generalization in the context of admissions testing. However, one study of the Law School Admission Test (LSAT) has established that about 70 percent of the apparent variation in LSAT validities among law schools is because of the statistical artifacts of sampling error, differences in test score ranges, and differences in criterion reliability.

While earlier professional standards encouraged local validation efforts, more recent guidelines have stressed the appropriateness of relying on the results of validity generalization studies, if the situation in which test use is contemplated is similar to those for which validity generalization was conducted. It was clear that ample data existed for investigating the extent to which the validity of the SAT might generalize across colleges. Over a period of nearly 20 years the College Board's Validity Study Service has generated some 3,000 validity studies for about 700 colleges. This impressive accumulation of studies shows a considerable variation in SAT validity coefficients across colleges, but until now no attempt has been made to apply validity generalization methods to these data.

The study reported here asked whether or not the validity of the SAT is both higher and less variable across colleges than it appears to be. Of particular interest was the influence on validity coefficients of (a) statistical artifacts (sampling error, restriction in the range of test scores, and criterion unreliability), (b) examinee self-selection to institutions, and (c) certain other social factors influencing the sorting of secondary school students to colleges.

The results of the study revealed that the average validity of both the SAT-V and SAT-M was estimated to be higher both for all test takers and for groups of applicants than for test takers on whom validity studies were based. For both SAT-V and SAT-M the average validity was .50 for all SAT takers, .39–.40 for applicants, and .33–.34 for validity study groups. This pattern of results suggested that, because of both institutional selection based on SAT scores and examinee self-selection in applying to particular schools, College Board validity studies underestimate the true validity of the SAT.

A significant proportion—but far from all—of the variation in observed validities was explained by sampling error and range restriction effects (53 percent for SAT-V and 45 percent for SAT-M). Another significant percentage would be accounted for by differences in criterion reliability.

Although validities could not be considered to be strictly equal across institutions, the ratio of the validity of SAT-V to that of SAT-M was nearly the same across colleges, suggesting that the factors associated with institutional uniqueness, whatever they may be, tend to operate about equally on SAT-V and SAT-M validity coefficients. The study did not provide any clues as to what these factors might be, nor was it able to identify any types of institutions on which such factors might operate.

Another conclusion reached was that negative, or other very low SAT-validity coefficients, should be regarded with suspicion, since they are more likely to have arisen from using small samples in validity studies, restriction of range of test scores, or the unreliability of the criterion.

Donald E. Powers

ABSTRACT

This study, which focused on the validity of the SAT-V and SAT-M, used data from 99 validity studies that were conducted by the Validity Study Service of the College Board. In addition to test validities based on first-year college averages, which were calculated using institutional data, validities for each college were also estimated for two other groups—applicants for admission to the colleges, and all SAT takers. These last two estimates were based on range restriction theory.

Substantial validity generalization was found: the assumption that applicant pool validities were all equal, together with sampling variance and the effects of selection, accounted for 36 percent and 34 percent of the variation of the SAT-V and SAT-M validities, respectively. The hypothesis of equal validity in pools like those of all SAT takers, plus sampling variance and the effects of selection, accounted for 53 percent and 33 percent of the variation of the SAT-verbal and SAT-mathematical validities, respectively. However, significant institutional uniqueness remains, though part of that uniqueness consists of variation in the reliability of first-year college average.

For these data, substantial validity was the rule. The average validities were quite high, rising to .55 for either SAT-V or SAT-M true scores for all SAT takers, and 95 percent of the observed validities were above .13 for SAT-V and .10 for SAT-M. Values below these may be owing to accidents of sampling, computing errors, or criterion defects,

and it should be noted that 95 percent is a conservative standard. Studies with slightly higher validities may be questioned as well, perhaps repeated, and the criterion examined carefully.

A hypothesis that validities for SAT-V and SAT-M might differ across institutions but have the same ratio was also tested. It was thought that departures from this second assumption might lead to the detection of institutional types. However, significant departures from the equal-ratio hypothesis tested at the 5 percent level occurred for about 5 percent of the institutions, so no detection of institutional types occurred.

INTRODUCTION

Traditionally, research on admissions testing has emphasized the results of local validity studies, that is, separate studies using data only from individual institutions, without regard to data collected at other, possibly similar, institutions. This practice, reinforced by the variation in test validities from institution to institution, has been regarded as consistent with professional standards for test use, which have embraced the notion that success may indeed be more predictable at some institutions than at others. The assumption was made that validity differences might arise from the unique characteristics of institutions or of the applicants they attract. These beliefs were also widely held in industrial applications of testing, in which test validity was thought to be highly specific to particular situations. For example, as late as 1975 the American Psychological Association's Division 14 (Industrial and Organizational Psychology) stated in its *Principles for the Validation and Use of Personnel Selection Procedures* (American Psychological Association 1975) that:

> Validity coefficients are obtained in specific situations. They apply only to those situations. A situation is defined by the characteristics of the samples of people, of settings, or criteria, etc. Careful job and situational analyses are needed to determine whether characteristics of the site of the original research and those of other sites are sufficiently similar to make the inference of generalizability reasonable.

An even more extreme view was espoused by the Equal Employment Opportunity Commission's (EEOC) *Guidelines on Employee Selection Procedures*, which required *every* use of an employment test to be validated (Equal Employment Opportunity Commission, et al. 1978).

However, research on institutional differences in test validity (Schmidt and Hunter 1977; Schmidt, Hunter, Pearlman, and Shane 1979; Pearlman, Schmidt, and Hunter 1980; Schmidt, Gast-Rosenberg, and Hunter 1980) led increasingly to the awareness that the effects of numerous presumed-to-be important variables were far less than supposed. In fact, much of the observed variation in test validity could be explained by statistical artifacts, most notably error resulting from the use of small samples and differences among institutions in (a) the distribution of test scores and (b) the reliability of the criterion. This growing awareness was reflected in the 1980 version of the Division 14 *Principles* (American Psychological Association 1980) as follows:

> Classic psychometric teaching has long held that validity is specific to the research study and that inability to generalize is one of the most serious shortcomings of selection psychology (Guion 1976). [But]...current research is showing that the differential effects of numerous variables may not be as great as heretofore assumed. To these findings are being added theoretical formulations, buttressed by empirical data, which propose that much of the difference in observed outcomes of validation research is due to statistical artifacts.... Continued evidence in this direction should enable further extensions of validity generalization.

In addition to acceptance by Division 14 of validity evidence from generalization studies in the personnel sphere, more general acceptance has been won. American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) have approved a revised edition of *Standards for Educational and Psychological Testing* (AERA, APA, NCME 1985). Principle 1.16 in these *Standards* states:

> When adequate local validation evidence is not available, criterion-related evidence of validity for a specified test use may be based on validity generalization from a set of prior studies, provided that the specified test-use situation can be considered to have been drawn from the same population of situations on which validity generalization was conducted.

Because the 1985 *Standards* cited above have become the approved policy of the professional organizations, validity generalization research is now accepted for a broad range of psychological applications, including education, by the relevant professional organizations in psychology and education. Thus, we see a trend away from site-specific interpretation of validity studies toward generalization across sites, attended by a degree of healthy uncertainty as to how far such generalization can be extended. In view of the generalizability of validities of tests in industrial settings, one might expect to find the same kind of result for admissions tests in academic settings.

The study reported here asked whether or not, after correcting for selection, self-selection, and the operation of other social factors, the validity of Scholastic Aptitude Test scores is similar across institutions, or at least generally substantial and positive. As mentioned above, a reasonable expectation of successful generalization was drawn from the results of Linn, Harnish, and Dunbar (1981), who conducted a validity generalization study using results from 726 law school validity studies (Schrader 1976). They found that statistical artifacts accounted for approximately 70 percent of the variance in validity coefficients. These authors did

not have data available for taking into account the use of previous grade-point average in admissions decisions, so the 70 percent figure might change, given a more complete modeling of admissions procedures. Indeed, none of the recent validity generalization studies have taken into account the use of multiple variables in admission, or hiring. The study reported here considered three preadmission variables: SAT-V, SAT-M, and the secondary school grade-point average.

ASSUMPTIONS AND HYPOTHESES

Because validity generalization research is in part concerned with the effects of selection on apparent test validity, the present study considered groups of examinees at three stages of selection. The first level was that of "all test takers," those who take the SAT during a given period of time. This group served as a standard population on which selection had not yet operated. The second level was that of "applicant pools," which consisted of SAT examinees who had applied to the sample of schools included in the study, and who differ from "all test takers" by virtue of the various social forces that influence their application behaviors. (The effect of these forces on the distribution of true test scores is discussed in Appendix B.) The third level consisted of examinees who were admitted to schools for which validity studies were conducted (College Board 1982). Having been (a) previously sorted into applicant groups, (b) selected by institutions, and (c) persistent in completing the first year of college, these "VSS groups" are therefore the most highly selected of those considered here.

This study tested the following generalization hypotheses in groups at each of the three levels of selection mentioned above:

1. that test validities are the same for all institutions,
2. that, although the validities may not be the same for all institutions, the *ratio* of the validity of SAT-V to that of SAT-M is the same.

The second hypothesis allowed for variation in criterion reliability among institutions.

It was not possible to conduct an ideal test of the hypotheses by randomly selecting and enrolling candidates from applicant pools or from all SAT-takers in order to calculate validities. But such hypothetical validities could be estimated by modeling the admissions process, and the processes by which all SAT-takers are sorted into applicant pools. These processes did not need to be modeled in their entirety because we were interested only in estimating correlations. Two models were used, that of classical test theory and K. Pearson's range restriction formulas (Gulliksen 1950) (Lord 1968). (Classical test theory provides useful theorems dealing with errors of measurement and true scores, and range restriction theory deals directly with adjusting covariances for the effects of selection.) The assumptions involved concern linearity and homoscedasticity in certain regressions. To specify the regressions for which the assumptions were needed, we distinguished between variables on which selection is made directly (explicit selectors), and other variables whose distributions are only indirectly affected by the selection. Assumptions of linearity and homoscedasticity were made for the regressions of the variables subject to selection on the explicit selectors. It remains to identify the variables that will take these roles.

Pearson's formulas have an unique advantage for this study in that their use does not require any information about institutional selection policies, as long as the right variables are used. Thus, the same equations can be used for all institutions, even those that differ in the weight given to the tests and secondary school performance measures.

In addition, we assume that the VSS group for an institution was selected explicitly from the applicant pool on the basis of SAT-V, SAT-M, and the secondary school grade-point average or rank in class. These, then, were the explicit selectors that accounted for the difference between applicant pool statistics and VSS-group statistics. The first-year college grade-point average is subject to implicit selection by virtue of its relationship to SAT scores and secondary school grades. And, as will be seen below, the self-reported grade-point average or rank in class obtained through the Student Descriptive Questionnaire also has a role in inferring *actual* grade-point average or rank in class statistics for applicants, for whom this information is not available.

The explicit selectors involved in the self-selection of applicants from all SAT takers are much more elusive than the explicit selectors used in admissions, and cannot be modeled directly. However, an appeal to classical test theory is helpful. In classical test theory, a test score consists of two components, a true score and an error of measurement. The latter is regarded as random and uninfluenced by societal forces, and it is uncorrelated with true score.[1] Hence the societal forces affect only the true score distributions. The variance of this error is not affected by the sorting of all SAT takers into applicant groups. According to classical test theory, we need only find the test reliability in a population for which the test variance is known to calculate the variance of the error of measurement. The reliability for all the SAT takers can be computed from the reliability estimates that are available from individual test administrations. Hence, we are able to use true score theory to represent the effects of elusive self-selection factors and other variables that route people from test taking to applicant pools.

[1]Actually, the distribution of errors of measurement is not homoscedastic over the full range of test scores (see, for example, Lord 1984), but in the present study, the emphasis is more on the total distributions including the dense portions where variation in the standard error of measurement is the smallest per unit change in true score. I believe that the distortions due to heteroscedasticity are minimal.

PROCEDURES

Variables

The variables in this study were SAT-verbal (SAT-V) and SAT-mathematical (SAT-M) test scores, secondary school grade-point average, secondary school rank in class, self-reported grade-point average, self-reported rank in class, and first-year college average. The study was concerned primarily with the relationships between the test scores and the first-year average; the other variables were needed only for estimation purposes. Secondary school rank in class and grade-point average were needed to model the effects of selection because they are used explicitly in the admissions process.

Self-reported secondary school performance was used for the following reasons. As classically written, range restriction computations require variance-covariance matrices for the explicit selectors in both the applicant and selected groups. However, the variance-covariance matrix of the variables subject to selection and their covariances with the explicit selectors need be known only for the selected group. This information is typically available in the hiring situation. However, in the present situation, applicants' secondary school grade-point averages are available only for applicants who are eventually included in a validity study. Therefore, it is necessary to estimate variance-covariance statistics for grade-point average in the applicant pool. The self-reported grade or rank was available for this purpose, having been supplied by most test takers when registering for the test. The differences between self-reported grade statistics in the selected and unselected pools were used to estimate statistics for actual school grades in the unselected groups.

The possibility of simply substituting the self-reported secondary school performances in place of the actual statistics in order to simplify calculations was considered. However, the relationship between self-reports and actual performances was not as high as expected. The average correlation of the self-reported with actual secondary school performance was only .73 for grades, and −.49 for rank. (The correlation for rank is negative since good performance receives a numerically small rank.) Therefore, a more complex procedure was required to estimate statistics for actual grades.

Samples

Two sources of College Board data were available. Test analyses contained statistical data describing examinees from particular administrations. For this project, data from several administrations were combined to estimate SAT standard deviations, intercorrelations, and reliabilities for all SAT takers. The Student History File contains records of candidates' progression through several stages of the admissions process, as well as validity study data for those examinees who ultimately attended institutions that conducted studies. SAT scores available with the validity study data are from the candidates' most recent administration. A high school rank or grade was also available, as were responses to the Student Descriptive Questionnaire (SDQ). Only students having a complete set of SDQ, SAT, secondary school performance, and first-year college performance data were included in VSS groups. Of those whose data were used by the Validity Study Service for institutions selected for this study, 49,578 or 92 percent had complete data. The smallest VSS group contained 66 cases and the largest contained 3,619. The frequency distribution of first-year averages was examined for outliers and a very few cases with extremely disparate averages were eliminated because of their unduly strong influence on least squares regression.

For each institution included in the study, the Student History File was searched for all test takers who had complete test scores and SDQ data and whose scores were sent to that institution. There were 81,793 such cases. Those for a particular institution are referred to here as the institution's applicant pool. In range restriction terms, this is the "unselected group," for whom, as noted above, secondary school performance measures were not available, and for whom the use of self-reported secondary school performances was therefore required.

The institutions included in the study were selected at random, subject to their having complete data for at least 50 examinees. Half of the institutions used rank-in-class as the index of secondary school performance; the other half used a grade-point average. All studies were based on freshman classes entering in 1980 or 1981. No school was used more than once.[2] One hundred schools were selected, but one was subsequently dropped because the standard deviation of self-reported rank-in-class was much smaller for the applicant pool than for the VSS group. When used in range restriction computations, these numbers led to a negative test variance.

Analyses

The following stages were used to estimate test validities for the applicant pools and for all SAT takers: (1) examination of the relationship between self-reported secondary school performances and actual performances, (2) estimation of statistics for all SAT takers, (3) estimation of secondary school performance statistics in the applicant pools, (4) estimation of college performance statistics in the applicant pools, (5) estimation of SAT validities for all SAT takers. Steps (6) through (8) consisted of estimating generalized validities at each of the three levels of selection for each hypothesis, and "reversing" the range restriction computations back to the VSS groups to provide "implied" validities, that is, validities that would be observed if the generalization hypothesis were true. Step (9) consisted of comparing the implied and actual validities to evaluate each generalization hypothesis.

[2] The sole exception was that midway through the study it was learned that two studies were inadvertently included for one institution. Instead of reanalyzing all data, both of these studies were retained.

Step 1. Self-Reported vs. Actual Secondary School Performances

As noted previously, if these variables had been sufficiently highly correlated, self-reports could have served as proxies for the missing grade-point averages and ranks in class, thus simplifying the estimation procedures. But, as was explained, the correlations were not large enough, and Step 3 was implemented instead.

Step 2. All SAT Takers

Test analyses were available for the 14 administrations of the SAT in calendar years 1980 and 1981. These analyses contained sample sizes, and scaled score means, variances, correlations, and reliabilities of SAT-V and SAT-M. The within-administration statistics were used to estimate statistics for the group of all candidates in the two testing years. Since the reliabilities were available, the variance of the error of measurement could be computed for an administration as the test variance times (1—reliability). A weighted average of these figures was used as the variance of the error of measurement in the total test-taking population.

Step 3. Secondary School Performance Statistics in the Applicant Pool

With the exception of data on the secondary school performance of applicants, data on all preadmission variables were available for each VSS group and for each applicant pool. Thus, data were available for two explicit selectors (SAT-V and SAT-M) in both groups, for one explicit selector (secondary school performance) in only the restricted (VSS) group, and for a variable subject to selection (self-reported secondary school performance) in both groups. As was pointed out earlier, although this is not the usual configuration of data available for range restriction computations, it is sufficient for estimating the variance of the secondary school performance variable and its covariances with SAT scores for the applicant pools. (See Appendix C for formulas.)

Step 4. First Year Average Statistics in the Applicant Pool

Step 3 resulted in, for both the applicant and VSS groups, variance-covariance matrices of explicit selectors, SAT scores, and secondary school performance. Data on first-year average was available only in the VSS sample. This configuration of data availability, typical in projects that involve correcting for the effects of selection, allows the use of standard formulas to estimate the applicant pool statistics.

Step 5. SAT Validities for all SAT Takers

The applicant pools were regarded as the restricted groups that were selected from all SAT takers. *For each institution*, the validities for all SAT takers were computed. In doing so, SAT true scores took the role of explicit selectors, with SAT scores and first-year averages being subject to the effects of selection. Because the test statistics for all SAT takers were computed in Step 2, one needs only to correct the first-year average statistics. For this purpose we have the same configuration of information as in Step 4—explicit selector data are available in both pools, but data on the variable subject to selection is present only in the restricted groups. The correction formulas are the same as for Step 4, though with different variables playing the roles. Then, because the covariances with true scores are, according to test theory, the same as the covariances with actual test scores, and because the test score statistics for all SAT takers were known, validities for that group could be computed. After this step, covariances and correlations were available for all groups at all levels of selection.

Step 6. Generalization Based on VSS Groups

Each hypothesis was evaluated for each group by comparing implied VSS-group validities with observed validities. The implied validities were obtained by computing a simplified set of validities for the groups in which the generalization was made, and correcting for the effects of selection to obtain the VSS-pool statistics for the particular generalization. For the VSS groups, the hypothesis of equal validities was implemented by using the average validity for a test as its implied validity for each institution. The equal ratio hypothesis was tested with the formula given in Appendix D, which multiplies the average validities by a different constant for each institution and uses the result as a different set of implied validities.

Step 7. Generalization Based on Applicant Pools

For this step the test validities averaged over applicant pools were taken as theoretical applicant pool validities and, using the formula of Appendix E, "reverse" corrected for the effects of selection on SAT-V, SAT-M, and the secondary school performance variable to obtain the validities implied by the equal validity hypothesis. Another set of implied validities was obtained for the equal ratio hypothesis applying the ratio-preserving procedures of the formula of Appendix D to the applicant pool validities and correcting for the effects of selection.

Step 8. Generalization Based on All SAT Takers

For this step the test validities estimated for all SAT takers were averaged across groups and the averages taken as theoretical applicant pool validities. Then, using the formula of Appendix E, these theoretical validities were corrected first for selection on true test scores and then for selection on SAT-V, SAT-M, and the secondary school performance variable to obtain implied validities that should be observed in the VSS groups if the generalization hypothesis were true. Another set of implied validities was obtained by applying the ratio-preserving procedures of Appendix D to the validities for the pool of all SAT takers and correcting for the effects of selection.

Step 9. Evaluation of Results

The results of these computations were evaluated in several ways. First, for each test-hypothesis-group combination, the means and standard deviations for the implied and observed validities were compared, and the correlations between implied and observed validities were computed.

Second, the percent of variance of the observed

validities that was accounted for by the implied validities and sampling error was calculated. The percent accounted for by the implied validities was simply the square of the correlation between implied and observed validities multiplied by one hundred. The percent of variation of the observed validities accounted for by sampling error was calculated by averaging the error variances of the individual coefficients, dividing the average by the test variance, and multiplying by one hundred. The sample variances of the observed validities were calculated using the same formula used by Perlman et al. (1980).

Third, for the equal ratio hypothesis applied in the pool of all SAT takers, the differences between the actual and observed validities were tested for significance.

Fourth, the distribution of criterion reliabilities adopted by Linn et al. (1981) was used to test the assumption of equal true score validities across institutions. This was necessary because, even though we estimated each institution's validities in a standardized pool of all SAT takers, thus eliminating test reliability as a source of variation in validity coefficients, the true score validities were still affected by the criterion reliabilities. According to test theory, each observed validity equals the correlation between test and true score times the square roots of the reliabilities of both the test and the criterion. Therefore, assuming that the correlations between test and criterion true scores are the same at all institutions, then each observed validity is the product of the square root of the criterion reliability times a constant less than unity. The constant is less than one because it is the product of the true score validity and the square root of the test reliability, each of which is less than one. Under the hypothesis of equal true score validity then, the variance of the test validities is equal to the product of the squared constant and the variance of the square roots of the criterion reliabilities, which is less than the variance of the square roots of the criterion reliabilities alone. Thus the variance of the square roots of the criterion reliabilities overestimates the amount of validity variance that is accounted for by variation in those reliabilities. Linn et al. (1981) have proposed a plausible distribution of criterion reliabilities, which was used to compute the needed variance.

Finally, the means, standard deviations, correlations, and fifth percentile values of the validities were found for the VSS groups, the applicant pools, and all SAT takers. For the applicant pools and all SAT takers the statistics were obtained for both test and true score validities. Use was made of the fifth percentile because many validity generalization studies report a figure called the 95 percent credibility value, the value above which 95 percent of true score validities are expected over a series of studies.

RESULTS

The Test Analysis data used to generate statistics for all SAT takers is given in Table 1, which reveals that, except for the means, the statistics are rather stable over administrations. The reliabilities cluster closely around .9 and the variances are quite stable. The statistics for all SAT takers were, for SAT-V and SAT-M, respectively, as follows: means were 424 and 465, standard deviations were 108 and 114, standard errors of measurement were 32 and 34, and reliabilities were both .91. The correlation between SAT-V and SAT-M was .68.

Table 2 contains the means of the validities observed using VSS-group data, as well as those estimated under the six generalization hypotheses. Clearly, the generalization hypotheses all led to implied validities that were at the right level, on the average.

Table 3 contains the standard deviations of the validities observed using VSS-group data, as well as those estimated under each generalization hypothesis. Note that the standard deviations of the implied correlations based on the hypothesis of equal validities are greatly depressed, being highest for all SAT takers. The generalized validities based on the equal ratio hypothesis fit well regardless of the groups over which the generalization was made.

Table 4 contains the correlations of the validities observed using VSS-group data with those estimated under each generalization hypothesis for each group. Again, this table reveals that the equal validity generalization is best made for all SAT takers, and that the equal ratio generalization is very accurate for all groups.

Another figure of merit for evaluating the success of validity generalization is the percent of variance of observed validity coefficients accounted for by the implied validities and sampling variance. These figures appear in Table 5. It can be seen in Table 5 that in the pool of all SAT takers, the assumption of equal validities, range restriction, and sampling error accounted for 53 percent and 45 percent of the observed validity variance. The generalization of equal validities is best made for that group.

The percent of variance accounted for was not reported for the equal ratio hypothesis because the results of computations based on that hypothesis were much more subject to chance fluctuations than were those based on the equal validity hypothesis. This is so because the level of the generalized validities was greatly influenced by the institutional validities alone, and it is not apparent how to incorporate this effect into a computation of percent of variance accounted for. The computed results from the ratio hypothesis account for more than 100 percent of the variance, and this was not useful or reasonable.

The observed VSS validities were tested individually for significant differences from the generalized validities based on the ratio model applied to validities for all SAT takers. There being 99 coefficients, one would expect 5 percent, or almost five, of these tests to reject the null hypothesis. For SAT-V, four of them were significant, and for SAT-M five were significant. Since these results were at the chance level, no attempt was made to interpret the lack of fit of individual schools' validities as an indication of institutional types.

Table 6 contains the means, standard deviations, and fifth percentile values of test and true score validities for the VSS, applicant and SAT-taker groups. In it can be seen an expected increase in validity as one scans from the restricted VSS groups, through the applicants, to all SAT takers. Such an increase also occurs with the fifth percentile score. Note also that true score validities are greater than test score validities, but not greatly so. True score validities are not presented for the VSS groups because the test score reliabilities are not known in those groups. The table shows little difference in the standard deviations of validities.

The variance of the square roots of the reliabilities in the Linn distribution is 20 percent of the variance of the validities for SAT-V, and 17 percent of the variance of the validities for SAT-M. These percentages are overestimates as was pointed out above.

DISCUSSION

The context in which validity generalization research arose was that of industrial hiring. Substantial degrees of validity generalization have been reported in this context, that is, differences in observed validities have been accounted for by statistical artifacts such as restriction of the tests because of their use in hiring, and variation in criterion reliability. Occupations over which generalization has been made cover a wide a variety of settings, perhaps an even wider variety of settings than would be encountered across academic institutions, over which one might therefore also expect validity to generalize. This surmise is supported by Linn et al. (1981), who found 70 percent generalization in a study of law school validities. An expectation of the present study was that an even higher percentage of variance might be explained if a more complete modeling of the selection procedure were possible using range restriction techniques. Therefore, multivariate corrections for restriction on SAT scores, SAT true scores, and secondary school performance were employed. The data were more complete than has usually been the case in such studies, in that data on the actual applicant pools were available. In addition, we were able to construct a standardized national population to control the variation in test reliability among groups of applicants. Even so, the generalization hypothesis of equal validities accounted for only 53 percent of SAT-V, and 45 percent of SAT-M.

However, these percentages do not take into account the variation in criterion reliability. To do so, the variance in criterion reliabilities was computed using the approach outlined in the fourth point of Step 9 in the Analysis section above. The magnitude of the resulting figure was, for all SAT takers, 20 percent of the variance in SAT-V validities, and 17 percent of the variance in SAT-M validities for all SAT takers. Adding these percentages to those accounted for by the hypothesis of equal validity and by sampling error brings the total percentage to 73 for SAT-V and 50 for SAT-M. To add these percentages is not strictly correct because they are generated on different base distributions. But if the result is approximately correct, for SAT-V it agrees with the estimate for the LSAT by Linn et al. (1981). The result for SAT-M is slightly smaller. It is concluded that though the equal validity hypothesis accounts for a large portion of the variation of validities, there are nevertheless substantial differences among institutions. The excellent fit of theoretical correlations that were based on the equal ratio hypothesis leads one to conclude, however, that these differences are primarily in the level of validity, rather than the pattern.

The existence of real variation in the level of validity coefficients, whether in the applicant pool or the pool of all SAT takers, does not mean that nothing can reasonably be anticipated about validities. First, for no institution was a test validity zero or negative when calculated for the population of all SAT takers. In fact, no test or true score validity below zero was observed for any institution or group. Second, the average validities reported in Table 5 are substantial. Third, 95 percent of the true score validities for all SAT takers were above the values of .32 for SAT-verbal and .26 for SAT-mathematical, and at the VSS level the corresponding figures are .13 and .10, respectively. Therefore, very low or negative coefficients are much more likely to result from sampling error, extreme selectivity on the test, and criterion unreliability, than from some special characteristic of the institution. This result emphasizes a principle that should be more generally appreciated than it is: low validity coefficients for selected incumbents are, in themselves, insufficient evidence that a test is invalid, and may be poor estimates of the actual test validity.

REFERENCES

AERA, APA, NCME. 1985. *Standards for Educational and Psychological Testing*. Washington: American Psychological Association.

American Psychological Association, Division of Industrial Organizational Psychology. 1975. *Principles for the Validation and Use of Personnel Selection Procedures*. Dayton, Ohio: American Psychological Association.

American Psychological Association, Division of Industrial-Organizational Psychology. 1980. *Principles for the Validation and Use of Personnel Selection Procedures*. (Second Edition) Berkeley, California: American Psychological Association.

College Entrance Examination Board. 1982. *Guide to the College Board Validity Study Service*. New York: College Entrance Examination Board.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of

Justice. 1978. Adoption by Four Agencies of Uniform Guidelines on Employee Selection Procedures. *Federal Register* 43, 38290–12008.

Guion, R. M. 1976. Recruiting, Selection and Job Placement. In M. D. Dunnette, ed. *Handbook of Industrial and Organizational Psychology*. Chicago: Rand McNally.

Gulliksen, H. 1950. *Theory of Mental Tests*. New York: John Wiley & Sons.

Linn, R. L., D. L. Harnish, and S. B. Dunbar. 1981. Validity Generalization and Situational Specificity: An Analysis of the Prediction of First-year Grades in Law School. *Applied Psychological Measurement* 5, 281–289.

Lord, F. M., and M. R. Novick. 1968. *Statistical Theories of Mental Test Scores*. Reading, Massachusetts: Addison-Wesley.

Lord, F. M. 1984. Standard Errors of Measurement at Different Ability Levels. *Journal of Educational Measurement* 21, 239–243.

Pearlman, K., F. L. Schmidt, and J. E. Hunter. 1980. Validity Generalization Results for Tests Used to Predict Job Proficiency and Training Success in Clerical Occupations. *Journal of Applied Psychology* 65, 373–406.

Schmidt, F. L., I. Gast-Rosenberg, and J. E. Hunter. 1980. Validity Generalization Results for Computer Programmers. *Journal of Applied Psychology* 65, 643–661.

Schmidt, F. L., and J. E. Hunter. 1977. Development of a General Solution to the Problem of Validity Generalization. *Journal of Applied Psychology* 62, 529-540.

Schmidt, F. L., J. E. Hunter, K. Pearlman, and G. S. Shane. 1979. Further Tests of the Schmidt-Hunter Bayesian Validity Generalization Procedure. *Personnel Psychology* 32, 257-281.

Schrader, W. B. 1976. Summary of Law School Validity Studies, 1948–1975. *Law School Admissions Council Report No. 76–8*. Newtown, Pennsylvania: Law School Admission Service.

APPENDIXES

Appendix A. Tables

Table 1. Means, Standard Deviations, Reliabilities, Intercorrelations, and Number of Cases for SAT Administrations in 1979 and 1980

	Means		*Standard Deviations*		*Reliabilities*			
Admin.	*V*	*M*	*V*	*M*	*V*	*M*	*Corr. V & M*	*Number Cases*
1979								
Jan.	404	442	109	113	.92	.91	.70	113912
Mar.	423	457	110	113	.92	.91	.68	210486
May	432	478	107	113	.92	.91	.68	253354
June	433	470	107	113	.91	.91	.68	181435
Oct.	437	481	107	116	.92	.91	.67	129646
Nov.	434	476	107	114	.91	.91	.64	348954
Dec.	399	446	103	112	.91	.91	.68	204243
1980								
Jan.	400	442	107	115	.92	.91	.67	112961
Mar.	418	455	108	112	.92	.91	.66	179162
May	425	477	105	114	.90	.91	.68	214314
June	434	470	107	112	.91	.90	.66	156926
Oct.	436	477	107	115	.92	.91	.68	65156
Nov.	431	472	107	114	.92	.91	.70	303457
Dec.	409	450	104	113	.91	.91	.67	197727

Table 2. Mean Validities for the VSS Groups, Observed and Implied, Based on the Two-Generalization Hypotheses Applied in the Three Sets of Groups

	Verbal	*Math*
Observed in VSS Data	.34	.33
Implied Validities Obtained Using Hypothesis and Group Indicated		
VSS Groups		
Equal Validities	.34	.33
Equal Ratios	.34	.33
Applicant Pool		
Equal Validities	.33	.33
Equal Ratios	.34	.33
All SAT Takers		
Equal Validities	.33	.32
Equal Ratios	.34	.33

Table 3. Standard Deviations of Validities for the VSS Groups, Observed and Implied, Based on the Two-Generalization Hypotheses Applied in the Three Sets of Groups

	Verbal	*Math*
Observed in VSS Data	.11	.11
Implied Validities Obtained Using Hypothesis and Group Indicated		
VSS Groups		
Equal Validities	.00	.00
Equal Ratios	.10	.10
Applicant Pool		
Equal Validities	.05	.06
Equal Ratios	.10	.11
SAT Takers		
Equal Validities	.07	.08
Equal Ratios	.10	.11

Table 4. VSS-Group Correlations Between the Observed and Implied Validities, Based on the Two-Generalization Hypotheses Applied in the Three Sets of Groups

Hypothesis	*Verbal*	*Math*
VSS Groups		
Equal Validities	.00	.00
Equal Ratios	.93	.93
Applicant Pool		
Equal Validities	.34	.32
Equal Ratios	.91	.91
SAT Takers		
Equal Validities	.52	.40
Equal Ratios	.89	.92

Table 5. Percentage of VSS-Group Validities Accounted for by Sampling Error and the Equal-Validity Hypothesis Applied in the Three Sets of Groups

Hypothesis	*Verbal*	*Math*
VSS Groups	26	29
Applicant Pool	38	39
SAT Takers	53	45

Table 6. Means, Standard Deviations, and Fifth Percentile Values of Test and True Score Validities for the VSS, Applicant, and SAT-Taker Groups

	Mean		*Standard Deviation*		*Fifth Percentile*	
	SAT-V	*SAT-M*	*SAT-V*	*SAT-M*	*SAT-V*	*SAT-M*
VSS Groups						
Test Score	.34	.33	.11	.11	.13	.10
Applicant Groups						
Test Score	.40	.39	.10	.10	.20	.15
True Score	.42	.42	.11	.11	.22	.16
SAT Takers						
Test Score	.50	.50	.10	.11	.30	.24
True Score	.52	.52	.11	.11	.32	.26

Appendix B. Use of Test Theory to Represent the Effects of Self-Selection

We make a very unrestrictive assumption that self-selection and external forces that steer a person toward a particular institution can be represented by a vector of variables, and it will be seen that they do not need to be identified. The variables can be represented by the vector variable X. Suppose that for all SAT takers the joint distribution of these variables with the true score T is a function J(X,T), and assume that errors of measurement are independent of X and T, with distribution D(E). Then, for all SAT takers, the joint distribution of all these variables is J(X,T)D(E), and the joint distribution of T and E would just be the marginal distribution of T times D(E).

Now suppose self-selection takes place. By hypothesis, and not a very restrictive one, it occurs by operation of explicit selection of X, and could be represented as G(X)J(X,T), where G adjusts the frequencies according to however the selection worked. Note that selection doesn't operate explicitly on T, since T cannot be observed. There would be a different G for each institution, and the marginal distribution of T for that institution would be the integral over the space of X of the product of G and J. Since the errors of measurement are independent by hypothesis, the distribution of E would be unaffected, but there would be an adjustment in the distribution of T. Hence the test score distributions would differ only by the distribution of T, with the distribution of E conditional on T being unaffected, and the range restriction formulas applying. Thus X operates on T so that even though T is not an explicit selector, it can take that role in the range restriction formulae because the conditional distributions of E are not affected. In particular, the standard error of measurement is unaffected by the selection, and because the expectation of errors of measurement given true score is zero, the covariance of test scores with true scores and the variance of true scores are equal in both the selected and unselected groups, hence the regression constants are the same in both groups.

Note, as was mentioned above, the really helpful fact that the variables in X need not be known, nor do the forms of J and G.

Appendix C. Use of a Supplementary Variable When Data Are Missing for an Explicit Selector

When capturing data for the routine operations of a secure testing program it is necessary only to obtain test scores and intended application information for the large bulk of candidates. Some candidates may attend an institution that will supply data to the program operator for use in a validity study in which the relationships of test scores, sending institution grades, and receiving institution grades of applicants are studied. If it is desired to estimate validity in an applicant pool, one needs statistics for the explicit selectors in both the applicant and incumbent pools in order to make the needed corrections for the effects of selection. One lacks, however, the sending institution statistics in the applicant pool, and they must therefore be estimated. This can be done if a supplementary variable exists that is present in both the applicant and incumbent pools, and that is correlated with the missing explicit selector. This supplementary variable takes the role of a variable subject to the effects of selection. The fact that it is observed in both the applicant and incumbent pools enables one to use it to estimate the missing statistics.

In the present case, the sending institutions are secondary schools, the receiving institutions are colleges, the incumbents are the college students whose data are used, the applicant pool are those who apply to the receiving institution, the explicit selectors that act on the applicant pool to create the incumbent pool are SAT-V, SAT-M, and a secondary school performance measure as a grade or rank in class. The supplementary variable can be a self-reported analog to the secondary school performance measure since a test program can easily collect candidate-reported biographical information along with the application to take the test.

The range restriction assumptions are that the coefficients of regression of variables subject to selection on the explicit selectors are undisturbed by the selection process, as are the errors of prediction of the variables subject to selection by the explicit selectors. Therefore the following normal equations for estimating regression coefficients in the applicant pool are satisfied by regression coefficients calculated in the incumbent pool.

$$C_{vs} = C_{vv} B_v + C_{vm} B_v + C_{vp} B_p \qquad (1)$$

$$C_{ms} = C_{vm} B_v + C_{mm} B_m + C_{mp} B_p \qquad (2)$$

In equations (1) and (2) all quantities are scalars. The subscripts v, m, p, and s stand for verbal, math, actual secondary school performance, and self-reported secondary school performance, respectively. C_{xy} is the covariance of x and y calculated in the applicant pool and Bx is the coefficient of partial regression of s on x calculated in the incumbent pool, hence known. Further, all covariances for which both variables are observed by the test program in the applicant pool are known. This leaves C_{vp} and C_{mp} as the only unknown quantities in equations (1) and (2) respectively. Therefore

$$C_{vp} = (C_{vs} - C_{vv} B_v - C_{vm} B_m)/B_p \qquad (3)$$

and

$$C_{mp} = (C_{ms} - C_{vm} B_v - C_{mm} B_m)/B_p \qquad (4)$$

From the assumption that the errors of prediction are unaffected by the selection process we obtain

$$C_{ss} = c_{ss} + b'(\|C_{xx}\| - \|c_{xx}\|)b \qquad (5)$$

where $\|C_{xx}\|$ and $\|c_{xx}\|$ are the explicit selector variance-covariance matrices in the applicant and incumbent pools, C_{ss} and c_{ss} are secondary school performance variances in those respective pools and b is a column vector of partial

regression coefficients, the B_x. b' is the transpose of b as in standard matrix notation. With the computations of (3) and (4) completed, all the quantities for (5) are available except the variance of the sending institution performance in the applicant pool, C_{pp}. If we define $||C^*_{xx}||$ as being the same as $||C_{xx}||$ but with a zero in the position of $||C_{pp}||$, which is the third column and third row, then (5) becomes

$$C_{ss} = c_{ss} + b'(||C^*_{xx}|| - ||c_{xx}||)b + C_{pp}B_\beta^2 \qquad (6)$$

All quantities in (6) are known except C_{pp}, for which the solution is

$$C_{pp} = (C_{ss} - c_{ss} - b'(||C^*_{xx}|| - ||c_{xx}||)b)/(B_\beta^2) \qquad (7)$$

With the calculation of C_{pp} in (7) all the entries $||C_{xx}||$ are available.

Appendix D. Generalizing the Assumption that the Validities Are Proportional Across Institutions

According to the hypothesis that the ratio of the validity of SAT-V to that of SAT-M is the same across institutions, the validity, V_{it} of a test, t, for institution i, is the product of F_t, a constant associated with the test, and G_i, a constant associated with the institution. Then

$$V_{it} = F_t G_i + \text{error} \qquad (1)$$

Neglecting the error,

$$V_{.t} = F_t G. \qquad (2)$$

where the dot indicates averaging over the missing variable, i in this case. Then, using (1) and (2)

$$V_{it}/V_{.t} = G_i/G. \qquad (3)$$

for either SAT-V or SAT-M. Therefore,

$$(V_{iv}/V_{.v} + V_{im}/V_{.m})/2 = G_i/G. = K_i \qquad (4)$$

Then from (2),

$$V_{.t}K_i = F_t G_i \qquad (5)$$

Thus (5) gives the formula for estimating the validity of the test under the hypothesis that the ratio of the validity of SAT-V to that of SAT-M is the same for all institutions. Since the average test validities are both multiplied by the same value, K_i, their ratio is the same for all groups, and their level varies with variation in the magnitude of K_i.

Appendix E. Calculating Validities in the Restricted Group

In the present study, once the generalization has been made, we want to reverse the range restriction calculations from the population of all SAT takers to an applicant pool, or from an applicant pool to a VSS group. In this situation, the validity in the unselected group is known, but not the validity in the restricted group, one would wish to have a restricted criterion variance that is consistent with the restricted validity. If we regard as our unknown the ratio of the criterion variances, there is enough information to solve the problem. This can be seen as follows. The assumption that selection does not affect the regression function can be written as follows:

$$C^{-1}_{xx}C_{xy} = c^{-1}_{xx}c_{xy} = b \qquad (1)$$

where C_{uv} and c_{uv} are covariance matrices for variables u and v for the unrestricted and restricted populations, respectively. The x variables are the explicit selectors, y is subject to selection and is a single variable in this project. Therefore the covariances involving both x and y are arranged in a column vector, as is b. Equation (1) can be rewritten as

$$MS_y = c^{-1}_{xx}s_x r_{xy} s_y, \qquad (2)$$

where

$$M = C^{-1}_{xx}S_x R_{xy}.$$

S_y and s_y are the standard deviations of y for the unrestricted and restricted populations respectively, S_x and s_x are diagonal matrices of standard deviations of the explicit selectors for the restricted and unrestricted populations, respectively, and R_{xy} and r_{xy} are the correlations between explicit selectors and y for the unselected and selected populations, arranged as column vectors. The assumption that selection does not affect the errors of prediction of y by x can be written as follows:

$$S^2_y = s^2_y + b'(C_{xx} - c_{xx})b$$

Hence

$$S^2_y = s^2_y + M'(C_{xx} - c_{xx})M S^2_y$$

Therefore,

$$S_y/s_y = 1/(1 - M'(C_{xx} - c_{xx})M)^{.5} = F \qquad (3)$$

Then using (2), and the definitions of M and F,

$$s^{-1}_x c_{xx} MF = r_{xy} \qquad (4)$$

All the information needed to carry out the calculation indicated on the left hand side of equation (4) is known after the generalized unrestricted validities are found.